On representing chemical environments

Albert P. Bartók*[a], Risi Kondor[b] and Gábor Csányi[a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

We review some recently proposed methods to represent atomic neighbourhood environments and analyse their relative merits in terms of their faithfulness and suitability for fitting potential energy surfaces (PES). The crucial properties that such representations (commonly called *descriptors*) must have are continuity and invariance to the basic symmetries of physics: rotation, reflection, translation, and permutation of identical atoms. We demonstrate that schemes that initially look quite different are specific cases of a general approach, in which a finite set of basis functions with increasing angular wave numbers are used to expand the atomic neighbourhood density function. We quantitatively show using the example system of small clusters that this expansion needs to be carried to higher and higher wave numbers as the number of neighbours increases in order to obtain a faithful representation, and that variants of the descriptors converge at very different rates.

1 Introduction

The appropriate representation of atomic environments in terms of a descriptor, a tuple of typically real valued functions of the atomic positions, is a crucial ingredient of algorithms used in modern computational chemistry and condensed matter physics. For example, in structure search applications[1], each configuration depends numerically on the precise initial conditions and the path of the search, so it is important to be able to identify equivalent structures and detect similarities. In molecular dynamics simulations of phase transitions[2], one needs good order parameters that are capable of detecting changes in the local order around given atoms. When constructing interatomic potentials and fitting potential energy surfaces (PES)[3–6], the driving application behind the present work, the functional forms depend on components of a carefully chosen representation of atomic neighbourhoods, *e.g.* bond lengths, bond angles, etc. "In silico" drug discovery[7,8] and other areas of chemical informatics also rely on characterising molecules using similar descriptors (also called fingerprints).

While specifying the position of each atom in a Cartesian coordinate system provides a simple and unequivocal description of atomic configurations, it is not directly suitable for making comparisons between structures: the list of coordinates is ordered arbitrarily and two structures might be mapped to each other by a rotation, reflection or translation so that two different lists of atomic coordinates can, in fact, represent the same or very similar structures. A good representation is *invariant* with respect to permutational, rotational, reflectional and translational symmetries, while retaining the faithfulness of the Cartesian representation. In particular, a system of invariant descriptors $q_1, q_2, \ldots, q_M$ is said to be *complete* if it uniquely determines the atomic environment, up to translation, rotation, and reflection. It is said to be *over-complete* if it contains spurious descriptors in the sense that a proper subset of $\{q_1, q_2, \ldots, q_M\}$ is, by itself, complete. If a representation is *complete*, a one-to-one mapping (a bijection) is obtained between the genuinely different atomic environments and the set of invariants comprising the representation. An *over-complete* representation assigns potentially many distinct descriptors to a given atomic structure, but guarantees that genuinely different atomic structures will never have identical descriptors associated with them: the function relating representations to atomic structures is a surjection.

Fitting potential energy surfaces (PESs) of small molecules to data generated by first principles electronic structure calculations has been a mainstay of computational chemistry for decades[9–23]. In these applications, the number of atoms which comprise the configuration space is relatively small, and, crucially, fixed. Typically, when modelling the PES of a small cluster, pairwise distances – or, equivalently, reciprocal distances – are used[4]. Such descriptions work only for a fixed number of atoms in a fixed order, thus they are limited to describing a given set of atoms. Even in this case, a seemingly new configuration is obtained by just permuting the order of atoms, i.e. crucial symmetries may be missing in this framework. Braams and Bowman[5] remedied this shortcoming by using polynomials of basis functions of the pairwise distances, selected so that they are invariant to the permutation of identical atoms. Computer code is available that generates the per-

[a] *Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom. Fax: +44 1223 332662; Tel: +44 1223 748536; E-mail: ab686@cam.ac.uk*
[b] *Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, IL 60637, United States of America.*

mutationally invariant polynomials automatically[5] (up to ten atoms), but this approach does not allow for varying number of atoms in the database of configurations.

In order to generate interatomic potentials for solids or large clusters capable of describing a wide variety of conditions, the number of neighbours that contribute to the energetics of an atom has to be allowed to vary, with the descriptor remaining continuous and differentiable. None of the traditional representations fulfil this criterion. Recently, however, a number of new, promising descriptors have been proposed together with potential energy surfaces based on them[3,6,24–26], and at this point it is not clear which method of representing atomic neighborhoods will prove to be optimal in the long term. In order to disentangle this issue from the rather complex details of generating first principles data and fitting PES, this paper focuses solely on the problem of descriptors.

The most well known invariants describing atomic neighbourhoods are the bond-order parameters originally proposed by Steinhardt et al.[27] These have been successfully used as order parameters in studies of nucleation[28], phase transitions[29] and glasses.[30] Here we show that the bond-order parameters form a subset of a more general set of invariants called the *bispectrum*[31]. The formally infinite array of bispectrum components provides an over-complete system of invariants, and by truncating it one obtains representations whose sensitivity can be refined at will. We relate the bispectrum to the representation proposed by Behler *et al.*[3,32], and show that, together with another descriptor set described below, their angular parts are all simple polynomials of the same canonical invariants.

2 Descriptors

Among the applications mentioned in the introduction, some require representing the geometry of an entire molecule, while for others one needs to describe the neighbourhood of an atom perhaps within a finite cutoff distance. While these two cases are closely related, the descriptors for one are not directly suitable for the other, although often the same idea can be used to derive representations for either case. Below, we focus on representing the neighbour environment of a single atom, but for some cases briefly mention easy generalisations that yield global molecular descriptors.

For N atomic position vectors $\{\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iN}\}$, taken relative to a central atom i, the symmetric matrix

$$\Sigma = \begin{bmatrix} \mathbf{r}_{i1}\cdot\mathbf{r}_{i1} & \mathbf{r}_{i1}\cdot\mathbf{r}_{i2} & \cdots & \mathbf{r}_{i1}\cdot\mathbf{r}_{iN} \\ \mathbf{r}_{i2}\cdot\mathbf{r}_{i1} & \mathbf{r}_{i2}\cdot\mathbf{r}_{i2} & \cdots & \mathbf{r}_{i2}\cdot\mathbf{r}_{iN} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{r}_{iN}\cdot\mathbf{r}_{i1} & \mathbf{r}_{iN}\cdot\mathbf{r}_{i2} & \cdots & \mathbf{r}_{iN}\cdot\mathbf{r}_{iN} \end{bmatrix} \tag{1}$$

is, according to Weyl,[33] an over-complete table of basic invariants to rotation, reflection and translation. However, Σ is not a suitable descriptor, because permutations of atoms change the order of rows and columns. For example, swapping atoms 1 and 2 results in the transformed matrix

$$\begin{bmatrix} \mathbf{r}_{i2}\cdot\mathbf{r}_{i2} & \mathbf{r}_{i2}\cdot\mathbf{r}_{i1} & \cdots & \mathbf{r}_{i2}\cdot\mathbf{r}_{iN} \\ \mathbf{r}_{i1}\cdot\mathbf{r}_{i2} & \mathbf{r}_{i1}\cdot\mathbf{r}_{i1} & \cdots & \mathbf{r}_{i1}\cdot\mathbf{r}_{iN} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{r}_{iN}\cdot\mathbf{r}_{i2} & \mathbf{r}_{iN}\cdot\mathbf{r}_{i1} & \cdots & \mathbf{r}_{iN}\cdot\mathbf{r}_{iN} \end{bmatrix}. \tag{2}$$

To compare two structures using their Weyl matrices Σ and Σ', we define a *reference* distance metric

$$d_{\mathrm{ref}} = \min_{\mathbf{P}} ||\Sigma - \mathbf{P}\Sigma'\mathbf{P}^T||, \tag{3}$$

where $\mathbf{P}$ is a permutation matrix and the minimisation is over all possible permutations. This metric is not differentiable at locations where the permutation that minimises (3) changes. It would also be intractable to calculate exactly for large numbers of atoms, but nevertheless we will use this metric to assess the faithfulness of other representations.

One way to generate permutationally invariant continuous functions of the Weyl matrix is to compute its eigenvalues, indeed, such a descriptor was recently used to fit the atomisation energies of a large set of molecules[26]. However, the list of eigenvalues is very far from being complete, since there are only N eigenvalues, whereas the dimensionality of the configuration space of N neighbours is $3N-3$ (after the rotational symmetries are removed). It is also unclear how to make the descriptors based on the eigenspectrum continuous and differentiable as the number of neighbours varies. Other, continuous invariants shown later in this paper are very closely related to the elements of Σ.

Another straightforward way to compare structures is based on pairing the atoms from each and finding the optimal rotation and reflection that brings the two structures into as close an alignment as possible. For each pair of structures $\{\mathbf{r}_{ij}\}_{j=1}^{N}$ and $\{\mathbf{r}'_{ij}\}_{j=1}^{N}$, it is possible to order the atoms according to their distance from the central atom i – or use other heuristics – and compute

$$\Delta(\hat{R}) = \sum_{j}^{N} |\mathbf{r}_{ij} - \hat{R}\mathbf{r}'_{ij}|^2,$$

where $\hat{R}$ represents an arbitrary rotation and define the distance between two configurations as

$$\Delta = \min_{\hat{R}} \Delta(\hat{R}). \tag{4}$$

This distance clearly has all the necessary invariances and completeness properties, but, like d_{ref}, it is not suitable for parametrising potential energy surfaces, because it is again not differentiable: the reordering procedure and the minimisation over rotations and reflections introduce cusps.

In the field of molecular informatics, one popular descriptor is based on the histogram of pairwise atomic distances[34], similar to Valle's crystal fingerprint[35]. We will not consider it here, because it is unsuitable for fitting PES, as it is clearly not complete: *e.g.* from 6 unordered distance values it is not necessarily possible to construct a unique tetrahedron, even though the number of degrees of freedom is also 6.

2.1 Bond-order parameters

As a first step in deriving a continuous invariant representation of atomic environments, we define the local atomic density function associated with atom i as

$$\rho_i(\mathbf{r}) = \sum_j \delta(\mathbf{r} - \mathbf{r}_{ij}), \tag{5}$$

where the index j runs over the neighbours of atom i and $\mathbf{r}_{ij}$ is the vector from atom i to atom j. In writing (5) we have assumed that only one atomic species is present, but generalising to multiple species is straightforward, e.g. by introducing different weights in the sum, corresponding to different species. Determining which neighbours to include in the summation can be done by using a simple binary valued, or a smooth real valued cutoff function of the interatomic distance, or via a more sophisticated procedure *e.g.* Voronoi analysis.[27] The local atomic density is already invariant to permuting neighbours, because changing the order of the atoms in the neighbour list only affects the order of the summation. In order to simplify the following derivation, for now we omit the information on the radial distance to the neighbours, but will show later how the radial information can be included. The local atomic density function can be expanded in terms of spherical harmonics (dropping the atomic index i for clarity):

$$\rho(\hat{\mathbf{r}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\hat{\mathbf{r}}), \tag{6}$$

where $\hat{\mathbf{r}}$ is the point on the sphere corresponding to the direction of the vector $\mathbf{r}$, thus $\rho(\hat{\mathbf{r}})$ is the projection of $\rho(\mathbf{r})$ onto the unit sphere.

The properties of functions defined on the unit sphere are related to the group theory of SO(3), the group of three dimensional rotations. Spherical harmonics form an orthonormal basis set for $L_2(S^2)$, the class of square integrable functions on the sphere:

$$\langle Y_{lm} | Y_{l'm'} \rangle = \delta_{ll'} \delta_{mm'},$$

where the inner product of functions f and g is defined as

$$\langle f | g \rangle = \int f^*(\hat{\mathbf{r}}) g(\hat{\mathbf{r}}) \, \mathrm{d}\Omega(\hat{\mathbf{r}}),$$

where the surface element $\mathrm{d}\Omega(\hat{\mathbf{r}})$ can be expressed in terms of the polar angles θ and ϕ as

$$\mathrm{d}\Omega(\hat{\mathbf{r}}) = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi,$$

and the coefficients c_{lm} are given by

$$c_{lm} = \langle \rho | Y_{lm} \rangle = \sum_j Y_{lm}(\hat{\mathbf{r}}_{ij}). \tag{7}$$

The quantities Q_{lm} introduced by Steinhardt *et al.*[27] are proportional to the coefficients c_{lm}. Averaging the coefficients for atom i provides the atomic order parameters

$$Q^i_{lm} = \frac{1}{N_i} \sum_j Y_{lm}(\hat{\mathbf{r}}_{ij}), \tag{8}$$

where N_i is the number of neighbours of atom i. Furthermore, averaging over atoms in the entire system gives a set of global order parameters

$$\bar{Q}_{lm} = \frac{1}{N_b} \sum_{ij} Y_{lm}(\hat{\mathbf{r}}_{ij}),$$

where N_b is the total number of atom pairs included in the summation. Both sets are invariant to permutations of atoms and translations, but still depend on the orientation of the reference frame. However, rotationally invariant combinations can be constructed as

$$Q^i_l = \left[\frac{4\pi}{2l+1} \sum_{m=-l}^{l} (Q^i_{lm})^* Q^i_{lm} \right]^{1/2} \text{and} \tag{9}$$

$$W^i_l = \sum_{m_1,m_2,m_3=-l}^{l} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} Q^i_{lm_1} Q^i_{lm_2} Q^i_{lm_2} \tag{10}$$

for atoms and

$$\bar{Q}_l = \left[\frac{4\pi}{2l+1} \sum_{m=-l}^{l} \bar{Q}^*_{lm} \bar{Q}_{lm} \right]^{1/2}$$

$$\bar{W}_l = \sum_{m_1,m_2,m_3=-l}^{l} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{Q}_{lm_1} \bar{Q}_{lm_2} \bar{Q}_{lm_2}$$

for the entire structure. The factor in parentheses is the Wigner-3jm symbol,[36] which is zero unless $m_1 + m_2 + m_3 = 0$.

The numbers Q^i_l and W^i_l are called second-order and third-order bond-order parameters, respectively. It is possible to normalise W^i_l such that it does not depend strongly on the number of neighbours:

$$\widehat{W}^i_l = \left[\sum_{m=-l}^{l} (Q^i_{lm})^* Q^i_{lm} \right]^{-3/2} W^i_l.$$

For symmetry reasons, only coefficients with $l \geq 4$ have non-zero values in environments with cubic symmetry and $l \geq 6$ for environments with icosahedral symmetry. Different values correspond to crystalline materials with different symmetry, while the global order parameters vanish in disordered

phases, such as liquids. Bond-order parameters were originally introduced for studying order in liquids and glasses[27], but were soon adopted for a wide range of applications. They have been used to study the free energy of clusters[37,38], melting of quantum solids[39], nucleation[40], as well as to serve as reaction coordinates in simulations of phase transitions[28] and also to generate interatomic potentials[41].

Using some basic concepts from representation theory, we now prove that the second-order bond-order parameters are indeed rotationally invariant, then we show a more general set of third order invariants[31], of which the Qs and the Ws are a subset. An arbitrary rotation $\hat{R}$ operating on a spherical harmonic function Y_{lm} transforms it into a linear combination of spherical harmonics with the same l index:

$$\hat{R}Y_{lm} = \sum_{m'=-l}^{l} D^l_{mm'}(\hat{R})Y_{lm'},$$

where the $\mathbf{D}^l(\hat{R})$ matrices are known as the Wigner matrices, which form the irreducible representations of the three dimensional rotation group, SO(3). The elements of the Wigner matrices are given by

$$D^l_{mm'}(\hat{R}) = \langle Y_{lm}|\hat{R}|Y_{lm'}\rangle. \tag{11}$$

It follows that the rotation operator $\hat{R}$ acts on the function ρ as

$$\begin{aligned}\hat{R}\rho &= \hat{R}\sum_{l=0}\sum_{m=-l}^{l} c_{lm}Y_{lm} = \sum_{l=0}\sum_{m=-l}^{l} c_{lm}\hat{R}Y_{lm}\\ &= \sum_{l=0}\sum_{m=-l}^{l}\sum_{m'=-l}^{l} c_{lm}D^l_{mm'}(R)Y_{lm'}\\ &\equiv \sum_{l=0}\sum_{m'=-l}^{l} c'_{lm}Y_{lm'},\end{aligned}$$

thus the column vector $\mathbf{c}_l$ of expansion coefficients transforms under rotation as

$$\mathbf{c}_l \rightarrow \mathbf{D}^l(\hat{R})\mathbf{c}_l.$$

Making use of the fact that rotations are unitary operations on $L_2(S^2)$, it is possible to show that the matrices $\mathbf{D}^l$ are unitary,

$$\mathbf{D}^{l\dagger}\mathbf{D}^l = \mathbf{I},$$

leading us to a set of rotational invariants,

$$p_l = \mathbf{c}_l^\dagger \mathbf{c}_l. \tag{12}$$

which we call the rotational power spectrum due to the analogy with the familiar power spectrum of ordinary Fourier analysis. The power spectrum remains the same under rotations:

$$p_l = \mathbf{c}_l^\dagger\mathbf{c}_l \rightarrow \left(\mathbf{c}_l^\dagger\mathbf{D}^{l\dagger}\right)\left(\mathbf{D}^l\mathbf{c}_l\right) = \mathbf{c}_l^\dagger\mathbf{c}_l.$$

We also note that the elements of $\mathbf{c}_l$ transform under reflection about the origin as

$$\mathbf{c}_l \rightarrow (-1)^l\,\mathbf{c}_l, \tag{13}$$

thus the power spectrum is invariant to this symmetry operation. A comparison with equations (7), (8) and (9) shows that the second-order bond-order parameters are related to the power spectrum via the simple scaling

$$Q_l = \left(\frac{4\pi}{2l+1}p_l\right)^{1/2}.$$

The power spectrum is clearly not a complete descriptor for a general function $f(\hat{\mathbf{r}})$ on the sphere, for example consider the two different functions

$$f_1 = Y_{22} + Y_{2-2} + Y_{33} + Y_{3-3}$$

and

$$f_2 = Y_{21} + Y_{2-1} + Y_{32} + Y_{3-2},$$

which both have the same power spectrum, $p_2 = 2$ and $p_3 = 2$. However, for the restricted class of functions which are sums of delta functions (such as the local atomic density ρ in equation (5)) with the radial information removed, numerical experiments suggest that the power spectrum might be overcomplete.

2.2 The bispectrum

We generalise the concept of the power spectrum to obtain a larger set of invariants via the coupling of the different angular momentum channels[31,42]. Let us consider the direct product $\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}$, which transforms under a rotation as

$$\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2} \rightarrow \left(\mathbf{D}^{l_l}\otimes\mathbf{D}^{l_2}\right)\left(\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}\right).$$

It follows from the representation theory of compact groups that the direct product of two irreducible representations of a compact group can be decomposed into a direct sum of irreducible representations of the same group[43]. In case of the SO(3) group, the direct product of two Wigner matrices can be decomposed into a direct sum of Wigner matrices,

$$D^{l_1}_{m_1m'_1}D^{l_2}_{m_2m'_2} = \sum_{l,m,m'} D^l_{mm'}\left(C^{l,l_1,l_2}_{m,m_1,m_2}\right)^* C^{l,l_1,l_2}_{m',m'_1,m'_2},$$

where $C^{l,l_1,l_2}_{m,m_1,m_2}$ denote the Clebsch-Gordan coefficients or, using more compact notation,

$$\mathbf{D}^{l_l}\otimes\mathbf{D}^{l_2} = \left(\mathbf{C}^{l_1,l_2}\right)^\dagger\left[\bigoplus_{l=|l_1-l_2|}^{l_1+l_2}\mathbf{D}^l\right]\mathbf{C}^{l_1,l_2}, \tag{14}$$

where $\mathbf{C}^{l_1,l_2}$ denote the matrices formed of the Clebsch-Gordan coefficients. These are themselves unitary, so the vector $\mathbf{C}^{l_1,l_2}\left(\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}\right)$ transforms as

$$\mathbf{C}^{l_1,l_2}\left(\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}\right)\rightarrow\left[\bigoplus_{l=|l_1-l_2|}^{l_1+l_2}\mathbf{D}^l\right]\mathbf{C}^{l_1,l_2}\left(\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}\right). \quad (15)$$

We now define $\mathbf{g}_{l_1,l_2,l}$ by

$$\bigoplus_{l=|l_1-l_2|}^{l_1+l_2}\mathbf{g}_{l_1,l_2,l}\equiv\mathbf{C}^{l_1,l_2}\left(\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}\right),$$

i.e. $\mathbf{g}_{l_1,l_2,l}$ is that part of the RHS in equation (15) which transforms under rotation as

$$\mathbf{g}_{l_1,l_2,l}\rightarrow\mathbf{D}^l\mathbf{g}_{l_1,l_2,l}.$$

Analogously to the power spectrum, we can now define the bispectrum as the collection of scalars

$$b_{l_1,l_2,l}=\mathbf{c}_l^\dagger\mathbf{g}_{l_1,l_2,l},$$

which are invariant to rotations:

$$b_{l_1,l_2,l}=\mathbf{c}_l^\dagger\mathbf{g}_{l_1,l_2,l}\rightarrow\left(\mathbf{c}_l\mathbf{D}^l\right)^\dagger\mathbf{D}^l\mathbf{g}_{l_1,l_2,l}=\mathbf{c}_l^\dagger\mathbf{g}_{l_1,l_2,l}.$$

It follows from equation (13) that those elements of the bispectrum where l_1+l_2+l is odd, change sign under reflection about the origin. If invariance to reflection is required, we take the absolute value of these components or omit them.

Rewriting the bispectrum formula as

$$b_{l_1,l_2,l}=\sum_{m=-l}^{l}\sum_{m_1=-l_1}^{l_1}\sum_{m_2=-l_2}^{l_2}c^*_{lm}C^{l,l_1,l_2}_{m,m_1,m_2}c_{l_1m_1}c_{l_2m_2}, \quad (16)$$

the similarity to the third-order bond-order parameters becomes apparent. Indeed, the Wigner $3jm$-symbols are related to the Clebsch-Gordan coefficients through

$$\begin{pmatrix}l_1 & l_2 & l_3\\ m_1 & m_2 & m_3\end{pmatrix}=\frac{(-1)^{l_1-l_2-m_3}}{\sqrt{2l_3+1}}C^{l_1,l_2,l_3}_{m_1,m_2,-m_3} \quad (17)$$

and by substituting the spherical harmonics identity $Y_{lm}=(-1)^mY^*_{l-m}$ in equation (8) it follows that

$$Q^i_{lm}=(-1)^m(Q^i_{lm})^*. \quad (18)$$

Substituting the identities (17) and (18) into the definition (10) we obtain

$$W^i_l=\frac{1}{\sqrt{2l+1}}\times\sum_{m_1,m_2,m_3=-l}^{l}(-1)^{-m}C^{l,l,l}_{m_1,m_2,-m_3}Q^i_{lm_1}Q^i_{lm_2}(-1)^m(Q^i_{lm_3})^*,$$

thus the third-order parameters W_l are seen to be proportional to the diagonal elements of the bispectrum, $b_{l,l,l}$. Noting that $Y_{00}\equiv 1$, the coefficient c_{00} is the number of neighbours N, and using $C^{l,0,l_2}_{m,0,m_2}=\delta_{l,l_2}\delta_{m,m_2}$, the bispectrum elements $l_1=0$, $l=l_2$ are identical to the previously introduced power spectrum components:

$$b_{l,0,l}=N_i\sum_{m=-l}^{l}\sum_{m_2=-l}^{l}c^*_{lm}\delta_{m,m_2}c_{lm_2}=$$
$$=N_i\sum_{m=-l}^{l}c^*_{lm}c_{lm}=N_ip_l,$$

therefore,

$$Q_l\propto\sqrt{p_l}\propto\sqrt{b_{l,0,l}}$$
$$W_l\propto b_{l,l,l}$$

The first few terms of the power spectrum and bispectrum for atom i with three neighbours are shown below, where θ_{ijk} is the angle between the bonds ij and ik and the sums are over the neighbours.

$$p_0=\frac{9}{4\pi}$$
$$p_1=\frac{3}{4\pi}\left(\sum_{jk}\cos\theta_{ijk}+3\right)$$
$$p_2=\frac{5}{4\pi}\left(\frac{3}{2}\sum_{jk}\cos^2\theta_{ijk}+6\right)$$
$$p_3=\frac{7}{4\pi}\left(\frac{5}{2}\sum_{jk}\cos^3\theta_{ijk}-\frac{3}{2}\sum_{jk}\cos\theta_{ijk}+3\right)$$
$$p_4=\frac{9}{16\pi}\left(\frac{35}{2}\sum_{jk}\cos^4\theta_{ijk}-15\sum_{jk}\cos^2\theta_{ijk}+13\right)$$
$$b_{112}=\sqrt{\frac{15}{128\pi^3}}\left(\frac{3}{4}\left(\sum_{jk}\cos\theta_{ijk}\right)^2+\frac{3}{2}\sum_{jk}\cos^2\theta_{ijk}+5\sum_{jk}\cos\theta_{ijk}\right)$$
$$b_{213}=\frac{150}{8}\sqrt{\frac{7}{\pi^3}}\left(\frac{5}{2}\sum_{jk}\cos^3\theta_{ijk}+\frac{5}{4}\sum_{jk}\cos^2\theta_{ijk}\sum_{jk}\cos\theta_{ijk}-\frac{1}{2}\left(\sum_{jk}\cos\theta_{ijk}\right)^2+4\sum_{jk}\cos^2\theta_{ijk}-2\sum_{jk}\cos\theta_{ijk}+18\right)$$

2.3 Radial basis

Thus far we neglected the distance of neighbouring atoms from the central atom by using the unit-sphere projection of

the atomic environment. One way to introduce radial information is to complement the spherical harmonics basis in equation (6) with radial basis functions g_n,[44]

$$\rho(\mathbf{r}) = \sum_n \sum_{l=0} \sum_{m=-l}^{l} c_{nlm} g_n(r) Y_{lm}(\hat{\mathbf{r}}).$$

If the set of radial basis functions is not orthonormal, i.e. $\langle g_n | g_m \rangle = S_{nm} \neq \delta_{nm}$, after obtaining the coefficients c'_{nlm} with

$$c'_{nlm} = \langle g_n Y_{lm} | \rho \rangle,$$

the elements c_{nlm} are given by

$$c_{nlm} = \sum_{n'} \left(S^{-1}\right)_{n'n} c'_{n'lm}.$$

In practice, when constructing the bispectrum, either c'_{nlm} or c_{nlm} can be used.

Rotational invariance must only apply globally, and not to each radial basis separately, therefore the angular momentum channels corresponding to different radial basis functions need to be coupled. So although extending equations (12) and (16) simply as

$$p_{n,l} = \sum_{m=-l}^{l} c^*_{nlm} c_{nlm}$$

$$b_{n,l_1,l_2,l} = \sum_{m=-l}^{l} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} c^*_{nlm} C^{l,l_1,l_2}_{m,m_1,m_2} c_{nl_1m_1} c_{nl_2m_2}$$

provides a set of invariants describing the three-dimensional neighbourhood of the atom, this can easily lead to a poor representation. However, if the radial basis functions do not sufficiently overlap, the different channels will still not be fully coupled, and the representation will have spurious quasi-invariances to rotating subsets of atoms at approximately the same distance, as illustrated in figure 1.

To avoid this, it is necessary to choose basis functions that are sensitive in a wide range of distances, although this may reduce the sensitivity of each radial basis channel, if the functions are varying very slowly. The fine-tuning of the basis set is rather arbitrary, and there is no guarantee that a choice exists that is optimal or even satisfactory for all systems of interest.

We suggest constructing radial functions from cubic and higher order polynomials, $\phi_\alpha(r) = (r_{\text{cut}} - r)^{\alpha+2}/N_\alpha$ for $\alpha = 1, 2, \ldots, n_{\text{max}}$, normalised on the range $(0, r_{\text{c}})$ using

$$N_\alpha = \sqrt{\int_0^{r_{\text{cut}}} (r_{\text{cut}} - r)^{2(\alpha+2)} \mathrm{d}r} = \sqrt{\frac{r_{\text{cut}}^{2\alpha+5}}{2\alpha+5}}.$$

The orthonormalised construction

$$g_n(r) = \sum_{\alpha=1}^{n_{\text{max}}} W_{n\alpha} \phi_\alpha(r) \tag{19}$$

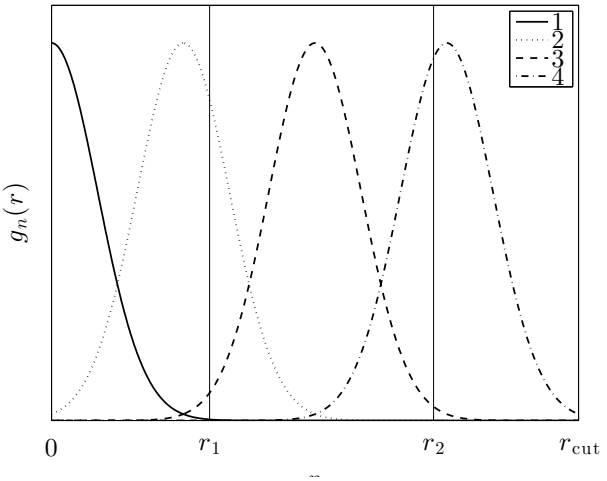

Fig. 1 Example of very localised radial basis functions, Gaussians in this case. Atoms 1 and 2 at distance r_1 and r_2 from the centre become decoupled as their contribution to the power spectrum or bispectrum elements is weighed down by the product $g_n(r_1)g_n(r_2)$, which is rather small for all n.

guarantees that each radial function returns smoothly to zero at the cutoff with both the first and the second derivative being zero. The linear combination coefficients are obtained from the overlap matrix as

$$S_{\alpha\beta} = \int_0^{r_{\text{cut}}} \phi_\alpha(r) \phi_\beta(r) \mathrm{d}r = \frac{\sqrt{(5+2\alpha)(5+2\beta)}}{5+\alpha+\beta}$$

$$\mathbf{W} = \mathbf{S}^{-1/2}.$$

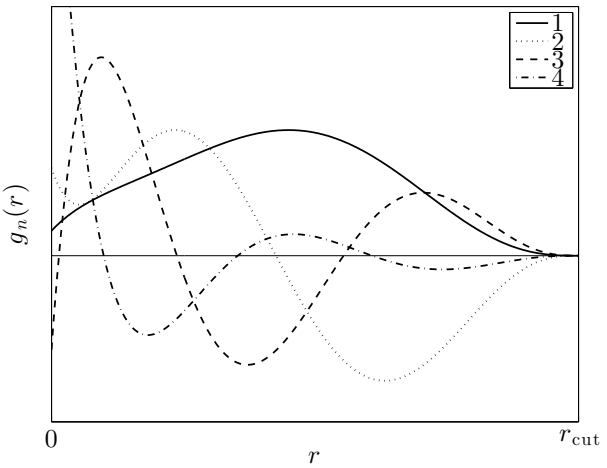

Fig. 2 Example of radial basis functions $g_n(r)$, as defined in equation (19) for $n = 1, 2, 3, 4$.

Another way to avoid radial decoupling is to define the rotational invariants in such a way that they couple different radial channels explicitly, for example, as

$$p_{n_1,n_2,l} = \sum_{m=-l}^{l} c^*_{n_1 lm} c_{n_2 lm} \text{ or}$$

$$b_{n_1,n_2,l_1,l_2,l} = \sum_{m=-l}^{l} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} c^*_{n_1 lm} C^{l,l_1,l_2}_{m,m_1,m_2} c_{n_2 l_1 m_1} c_{n_2 l_2 m_2}.$$

Here, each invariant has contributions from two different radial basis channels, and so we ensure that they cannot become decoupled, but at the price of increasing the number of invariants quadratically or even cubically in the number of radial basis functions.

2.4 4-dimensional power spectrum and bispectrum

We now present an alternative to the power spectrum and bispectrum that does not need the explicit introduction of a radial basis set, but still represents atomic neighbourhoods in three-dimensional space. We start with projecting the atomic neighbourhood density within a cutoff radius r_{cut} onto the surface of the four-dimensional sphere S^3 with radius r_0. The surface of S^3 is defined as the set of points $\mathbf{s} \in \mathbb{R}^4$, where $s_1^2 + s_2^2 + s_3^2 + s_4^2 = r_0^2$, while the polar angles ϕ, θ and θ_0 of $\mathbf{s}$ are defined so that

$$\begin{aligned} s_1 &= r_0 \cos\theta_0 \\ s_2 &= r_0 \sin\theta_0 \cos\theta \\ s_3 &= r_0 \sin\theta_0 \sin\theta \cos\phi \\ s_4 &= r_0 \sin\theta_0 \sin\theta \sin\phi. \end{aligned}$$

We choose to use the projection

$$\mathbf{r} \equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} \rightarrow \begin{aligned} \phi &= \arctan(y/x) \\ \theta &= \arccos(z/|\mathbf{r}|), \\ \theta_0 &= \pi|\mathbf{r}|/r_0 \end{aligned}$$

where $r_0 > r_{\text{cut}}$ is a parameter, thus rotations in three-dimensional space correspond to a subset of rotations in four-dimensional space. This projection is somewhat similar to a Riemann projection, except in a Riemann projection θ_0 would be defined as

$$\theta_0 = 2\arctan(|\mathbf{r}|/2r_0)$$

implying

$$\theta_0 \approx |\mathbf{r}|/r_0 \text{ for } |\mathbf{r}| \ll r_0.$$

In contrast to the Riemann projection, our choice of θ_0 allows more sensitive representation of the entire radial range. The limit $r_0 = r_{\text{cut}}$ projects each atom at the cutoff distance to the south pole of the 4-dimensional sphere, thus losing all angular information. Too large an r_0 would project all positions onto a small surface area of the sphere, requiring a large number of basis functions to represent the atomic environment. In practice, a large range of r_0 values works well, in particular, we used $r_0 = {}^4\!/_3\, r_{\text{cut}}$.

To illustrate the procedure, Figure 3 shows the Riemann projections for one and two dimensions, which can be easily drawn.

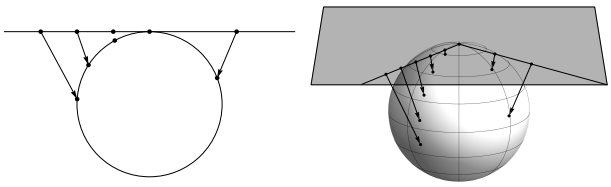

Fig. 3 Two- and three-dimensional Riemann constructions that map a flat space onto the surface of a sphere.

An arbitrary function ρ defined on the surface of a 4D sphere can be numerically represented using the hyperspherical harmonic functions $U^j_{m'm}(\phi,\theta,\theta_0)$ [36,45]:

$$\rho = \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} c^j_{m'm} U^j_{m'm}, \tag{21}$$

which, in fact, correspond to individual matrix components of the Wigner (i.e. rotational) matrices, as defined in equation (11). In this case the arguments represent a rotation by θ_0 around the vector pointing in the (ϕ,θ) direction, which can be transformed to the conventional Euler-angles, and j takes half-integer values.

The hyperspherical harmonics form an orthonormal basis, thus the expansion coefficients $c^j_{m'm}$ can be calculated via

$$c^j_{m'm} = \langle U^j_{m'm} | \rho \rangle,$$

where $\langle . | . \rangle$ denotes the inner product defined on the four-dimensional hypersphere:

$$\langle f|g\rangle = \int_0^{\pi} \mathrm{d}\theta_0 \sin^2\theta_0 \int_0^{\pi} \mathrm{d}\theta \sin\theta \int_0^{2\pi} \mathrm{d}\phi \; f^*(\theta_0,\theta,\phi)\, g(\theta_0,\theta,\phi).$$

Although the coefficients $c^j_{m'm}$ have two indices besides j, for each j it is convenient to collect them into a single vector $\mathbf{c}^j$. Similarly to the three-dimensional case, rotations act on the hyperspherical harmonic functions as

$$\hat{R} U^j_{m'_1 m_1} = \sum_{m'_2 m_2} R^j_{m'_1 m_1 m'_2 m_2} U^j_{m'_2 m_2},$$

where the matrix elements $R^j_{m'_1 m_1 m'_2 m_2}$ are given by

$$R^j_{m'_1 m_1 m'_2 m_2} = \langle U^j_{m'_1 m_1} | \hat{R} | U^j_{m'_2 m_2} \rangle.$$

Hence the rotation $\hat{R}$ acting on ρ transforms the coefficient vectors $\mathbf{c}^j$ according to

$$\mathbf{c}^j \to \mathbf{R}^j \mathbf{c}^j.$$

The unitary $\mathbf{R}^j$ matrices are the SO(4) analogues of the Wigner-matrices $\mathbf{D}^l$ of the SO(3) case above, and it can be shown that the direct product of the four-dimensional rotation matrices decomposes according to

$$\mathbf{R}^{j_l} \otimes \mathbf{R}^{j_2} = \left(\mathbf{H}^{j_1,j_2}\right)^\dagger \left[\bigoplus_{j=|j_1-j_2|}^{j_1+j_2} \mathbf{R}^j \right] \mathbf{H}^{j_1,j_2},$$

which is the 4-dimensional analogue of equation (14). The coupling constants $\mathbf{H}^{j_1,j_2}$, or Clebsch-Gordan coefficients of SO(4) are[45,46]

$$H^{jmm'}_{j_1 m_1 m'_1, j_2 m_2 m'_2} \equiv C^{j,j_1,j_2}_{m,m_1,m_2} C^{j,j_1,j_2}_{m',m'_1,m'_2}.$$

The rest of the derivation continues analogously to the 3D case, and finally we arrive at the expression for the SO(4) bispectrum elements

$$B_{j_1,j_2,j} = \sum_{m'_1,m_1=-j_1}^{j_1} \sum_{m'_2,m_2=-j_2}^{j_2} \sum_{m',m=-j}^{j} C^{j,j_1,j_2}_{m,m_1,m_2} C^{j,j_1,j_2}_{m',m'_1,m'_2} \times \\ \times \left(c^j_{m'm}\right)^* c^{j_1}_{m'_1 m_1} c^{j_2}_{m'_2 m_2},$$

and the 4D power spectrum can be constructed as

$$P_j = \sum_{m',m=-j}^{j} \left(c^j_{m'm}\right)^* c^j_{m'm}.$$

The 4D bispectrum is invariant to rotations of four-dimensional space, which include three-dimensional rotations. However, there are additional rotations, associated with the third polar angle θ_0, which, in our case, represents the radial information. In order to eliminate the unphysical invariance with respect to rotations along the third polar angle, we modify the atomic density as

$$\rho_i(\mathbf{r}) = \delta(\mathbf{0}) + \sum_j \delta(\mathbf{r} - \mathbf{r}_{ij}),$$

i.e. we add the central atom as a fixed reference point. The resulting invariants $B_{j_1,j_2,j}$ have only three indices, but contain both radial and angular information, and have the required symmetry properties. There are no adjustable parameters in the definition of the invariants, apart from the radial cutoff, r_{cut}, and the projection parameter r_0.

The number of components in the truncated representation depends on the band limit $j_{\max}$ in the expansion (21). For symmetry reasons, the bispectrum components with non-integer $j_1 + j_2 + j$ change sign under reflection and, because of this reason, we omitted them. Just as in the 3D case, the representation is probably over-complete, i.e. most of the bispectrum components are redundant. To reduce the number of redundant elements, we only used the 'diagonal' components, *i.e.* $j_1 = j_2$. Table 1 shows the number of bispectrum elements for increasing band limit values.

Table 1 Number of compontents in the full and diagonal bispectrum as a function of the band limit $j_{\max}$.

$j_{\max}$	0	½	1	3/2	2	5/2	3	7/2	4	9/2
$B_{j_1 j_1 j}$	1	2	5	7	12	15	22	26	35	40
$B_{j_1 j_2 j}$	1	4	11	23	42	69	106	154	215	290

2.5 Parrinello-Behler descriptor

We include in the tests below the descriptor suggested by Parrinello and Behler[3] using the parameters published recently[47] (and henceforth termed PB). The two- and three-body symmetry functions (in their terminology) are,

$$G^2_{i\alpha} = \sum_{j \neq i} = \exp\left[-\eta_\alpha (r_{ij} - R_{s\alpha})^2\right] f_c(r_{ij})$$

and

$$G^4_{i\alpha} = 2^{1-\zeta_\alpha} \sum_{j,k \neq i} (1 + \lambda_\alpha \cos\theta_{ijk})^{\zeta_\alpha} \exp\left[-\eta_\alpha (r^2_{ij} + r^2_{ik} + r^2_{jk})\right] \\ \times f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}),$$

where the cutoff function is defined as

$$f_c(r) = \begin{cases} \left[\cos\left(\frac{\pi r}{r_{\text{cut}}}\right) + 1\right]/2 & \text{for } r \leq r_{\text{cut}} \\ 0 & \text{for } r > r_{\text{cut}} \end{cases}.$$

Different values of the parameters $\eta, R_s, \zeta, \lambda$ can be used to generate an arbitrary number of invariants.

2.6 Angular Fourier Series

Notice that the angular part of the power spectrum, bispectrum (section 2.2) and the descriptors defined by Parrinello and Behler (section 2.5) are all simple polynomials of the canonical set $\sum_{jk} \cos^m \theta_{ijk}$ for integer m, which in turn are sums of powers of the basic invariants of Weyl. We include in the tests in the next section a further descriptor set, which we call the Angular Fourier Series (AFS) descriptor, formed

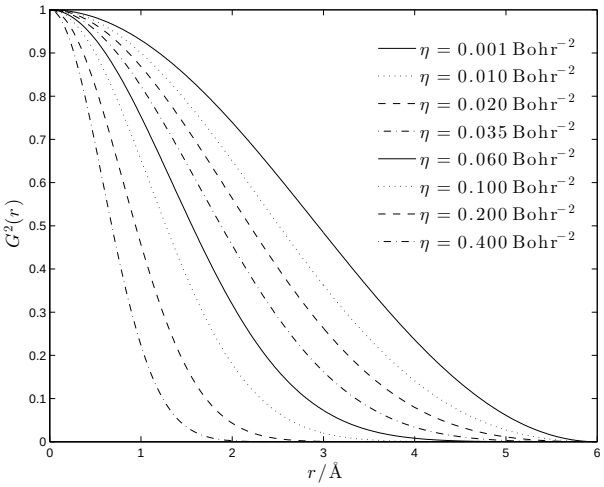

Fig. 4 Radial basis functions G^2 in the Parrinello-Behler (PB) type descriptors.[47]

by the orthogonal polynomials of the basic invariants, which are conveniently chosen as the Chebyshev-polynomials $T_l(x)$, as

$$T_l(\cos\theta) = \cos(l\theta),$$

and incorporate the radial information using the basis functions defined in equation (19), leading to

$$\mathrm{AFS}^{(i)}_{n,l} = \sum_{j,k} g_n(r_{ij}) g_n(r_{ik}) \cos(l\theta_{ijk}).$$

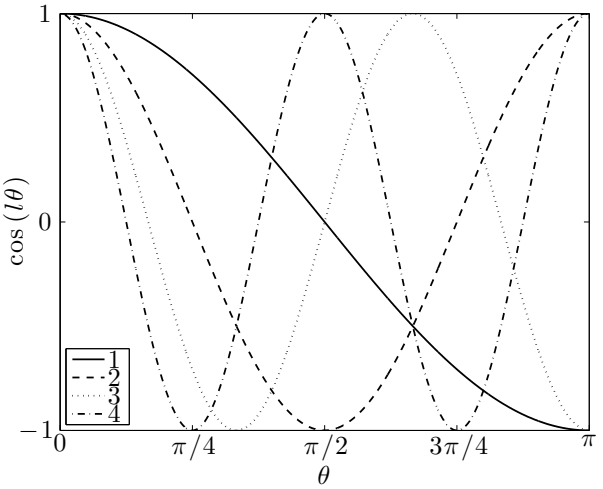

Fig. 5 Examples of the angular basis functions for $l_{\max} = 4$ of the AFS descriptor.

3 Reconstruction experiments

Recall that the elements $\mathbf{r}_{ij} \cdot \mathbf{r}_{ik}$ of Σ defined in equation (1) are an over-complete set of basic invariants, which, in the case of atoms scattered on the surface of a unit sphere ($|\mathbf{r}_{ij}| = 1$), are the cosines of the bond angles, θ_{ijk}. Thus, the angular part of the above descriptors are permutationally invariant functions of the basic invariants in Σ, and depending on the actual number of descriptor elements used, they may form a complete or over-complete representation of the atomic environment. There are $(N^2+N)/2$ basic invariants, thus we need at most this many independent descriptor elements, whereas the number of independent degrees of freedom in the neighbourhood configuration is $3N-3$, so it remains unclear how many of the descriptor elements are actually needed in order to reconstruct an atomic environment up to an arbitrary rotation and permutation of the N neighbour atoms. Obtaining exact results regarding completeness would be difficult, especially since the algebraically dependency relationships between the descriptor elements is unknown.

It is worth emphasising that severely incomplete descriptors will lead to poor potential energy surfaces, because entire manifolds of atomic configurations with widely varying true energies will be represented by the same descriptor and hence assigned the same energy in the fitted PES, resulting in many degenerate modes. As the mappings are highly nonlinear, in general we could not find an analytical route to invert the descriptor transformations and thus to analytically investigate the completeness properties. However, it is possible to conduct numerical experiments in which we compare the descriptors of a fixed target with that of a candidate structure and minimise the difference with respect to the atomic coordinates of the candidate. The purpose of these experiments is to determine if a representation is complete or not, and in the latter case to characterise the degree of its faithfulness.

The global minimum of the descriptor difference between the target and the candidate is the zero value and is always a manifold due to the symmetries built into the descriptors, but for an incomplete descriptor, many inequivalent structures will also appear equivalent, thus enlarging the dimensionality of the global minimum manifold. Further, it can be expected that the descriptor difference function has a number of local minima.

In our experiments we tried to recover a given target structure after randomising its atomic coordinates. For each n ($4 \le n \le 19$) we used 10 different Si_n clusters, obtained from a tight-binding[48] molecular dynamics trajectory at a temperature of 2000 K. For each cluster, we selected an atom as the origin, randomised its neighbours within the radial cutoff of the descriptors by some amount, and then tried to reconstruct the original structure by minimising the magnitude of the difference between the descriptors of the fixed target and the can-

didate as the atomic positions were varied. We repeated this procedure five times for each target configuration and then selected the reconstruction which ended up closest to the target. In some cases, such as large n and large randomisation, it was possible for our procedure to find only a local minimum during the minimisation, these were easily identified (by noting the small gradient of the objective function while the value of the objective function was not small) and discarded.

In order to assess the success of the reconstruction procedure we employed the reference measures defined in equations (3) and (4). However, in some cases it was difficult or impossible to find the right rotation $\hat{R}$ in (4), whereas d_{ref} in (3) proved reliable. For each d_{ref}, an initial $\mathbf{P}$ was generated by ordering the atoms according to their distances from the central atom, then the optimal permutation was found using a simple random search in the space of permutations.

We minimised the difference between the target and candidate descriptors in the space of atomic coordinates using the Conjugate Gradients algorithm, stopping the minimisation if either the gradient or the reference distance d_{ref} became smaller than 10^{-8} and 10^{-2}, respectively. In order to ensure that structures deemed non-equivalent by $d_{\mathrm{ref}} > 10^{-2}$ were genuinely different, we cross-checked them by noting the value of Δ from equation (4) and also employing the atomic fingerprints suggested by Oganov and Valle[35]. To give a sense of the typical magnitude of d_{ref} measure, the actual difference in terms of atomic distances between two example structures is shown in Figure 6.

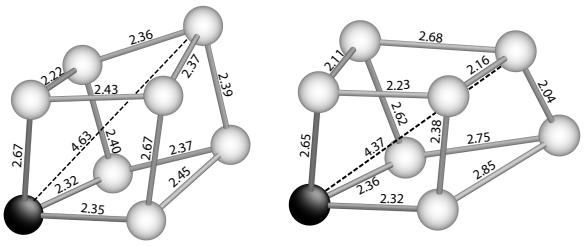

Fig. 6 Two Si_8 clusters that differ by $d_{\mathrm{ref}} = 4.1\text{Å}^2$. Atoms i and i' at the origin, *i.e.* the centre of rotations, is coloured black. In terms of Parinello-Behler type descriptors, the difference $\sum_\alpha (G_{i\alpha} - G_{i'\alpha})^2$ between the atomic environments of i (left) and i' (right) is $6 \cdot 10^{-7}$. The bond lengths are shown, in ångströms.

In the first set of recontruction experiments, in order to provide a fair comparison, the truncation of the formally infinite set of descriptors were chosen in such a way that the finite descriptors had roughly equal number of components: 51 in total for the bispectrum and PB descriptors and 50 for the AFS and power spectrum. This corresponds to a truncation of the 4-D bispectrum with $j_{\max} = 5$, the PB descriptor with its published parameters[47] and the AFS and power spectrum using $l_{\max} = 9$ and $n_{\max} = 5$. We note that in case of the PB descriptor the band limit of the angular descriptors (corresponding to our $l_{\max}$ or $j_{\max}$) was $\zeta_{\max} = 16$ and only the values $\zeta = 1, 2, 4, 16$ are used.

Figure 7 shows the quality of reconstruction based on the PB, AFS, 3D power spectrum, 3D bispectrum and 4D bispectrum as given by the average reference distance d_{ref} achieved. The general trend is the same for all descriptors: as the number of neighbours increases, the average d_{ref} increases, and thus the faithfullness of the recontruction decreases. Noting that the stopping criteron for the reconstruction process was $d_{\mathrm{ref}} < 10^{-2}$, larger randomisation of the initial atomic coordinates (bottom panel) reveals the poor representation power for all descriptors for $n > 10$, and the neighbour configuration becomes impossible to determine from the descriptor.

The poor quality of representation is partly attributable to the decrease in sensitivity to the positions of atoms near the cutoff. For example, Figure 8 shows two Si_8 clusters for which the all the descriptors failed to lead to perfect reconstructions (resulting in the observed peak on Figure 7). The atom marked A in the figure is within the 6 Å cutoff, but close to it. In order to separate out this effect, we repeated the reconstruction experiments with a radial cutoff of 9 Å (omitting the PB descriptor now since there is no published optimal parameter set for this cutoff). The results are shown in Figure 9 for the larger initial randomisation. The peak near $n = 8$ is now absent, and the transition from faithful reconstruction (for $n \leq 9$ for the 3D power spectrum and AFS, and for $n \leq 12$ for the 4D bispectrum) to failure for larger n is much clearer.

Since all the descriptors are likely to be over-complete when the infinite series of the basis set expansion is not truncated, the reconstruction quality is expected to increase with increasing descriptor length. To verify this, Figure 10 shows the reconstruction quality of the AFS descriptor for varying truncations of the angular part of the basis set. The representation becomes monotonically better for higher angular resolutions. However, this comes at the price of introducing ever more highly oscillating basis functions, which might be less and less suitable for fitting generally smooth potential energy surfaces.

To verify that the above results are not affected by artefacts of the minimisation procedure, *e.g.* getting stuck, Figure 11 shows the convergence of the reference measure d_{ref} during a minimisation as well as the corresponding convergence of the target function (the difference between the target and candidate descriptors). There was no difficulty in converging the target function to zero (the global minimum) for any of the descriptors, while the reference similarity converged to a nonzero value for incomplete descriptors.

4 Conclusion

In this paper we discussed a number of approaches to represent atomic environments such that the representation is invariant to global rotations, reflections, and permutations of atoms. We showed that the bond-order parameters introduced by Steinhardt et al. are equivalent to certain elements of the SO(3) power spectrum and bispectrum. To incorporate radial information and therefore provide a full description of the atomic environment we reviewed the construction of the SO(4) power spectrum and bispectrum as an alternative to introducing explicit radial basis functions. We also demonstrated that these constructs, as well as the descriptors suggested by Parrinello and Behler, in fact, use very similar terms and form part of a general family based on a Fourier series expansion of bond angles. In practice, when the Fourier series is truncated, the faithfulness of the descriptors decreases as the number of neighbours increases. The descriptors discussed in this paper are thus suitable for comparing atomic environments, with a tuneable tradeoff between the size (and thus computational cost) of evaluating the descriptor and its faithfulness in terms of its ability to represent the atomic environment uniquely up to symmetries. With typically used parameters, however, all descriptors failed for Si clusters with more than 13 atoms. The relative performance of the various representations in the context of fitting potential energy surfaces remains to be tested.

Acknowledgements

APB gratefully acknowledges funding from Magdalene College, Cambridge. Figures 6 and 8 were generated using AtomEye[49]. This work was partly supported by the European Framework Programme (FP7/2007-2013) under grant agreement no. 229205.

References

1 C. J. Pickard and R. J. Needs, *Nature Materials*, 2008, **7**, 775–779.
2 D. J. Wales, *Energy Landscapes*, Cambridge University Press, 2004.
3 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
4 L. Raff, R. Komanduri, M. Hagan and S. Bukkapatnam, *Neural Networks in Chemical Reaction Dynamics* , Oxford University Press, USA, 2012.
5 B. Braams and J. Bowman, *Int. Rev. in Phys. Chem.*, 2009, **28**, 577–606.
6 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.
7 O. Obrezanova, G. Csányi, J. M. R. Gola and M. D. Segall, *J. Chem. Inf. Model.*, 2007, **47**, 1847–1857.
8 M. Segall, *J. Comput. Aided Mol. Des.*, 2011, **26**, 121–124.
9 D. Tromans, *Hydrometallurgy*, 1998, **48**, 327–342.
10 J. Ischtwan and M. A. Collins, *J. Chem. Phys.*, 1994, **100**, 8080.
11 M. A. Collins, *Theor. Chem. Acc.*, 2002, **108**, 313–324.
12 T.-S. Ho, T. Hollebeek, H. Rabitz, L. B. Harding and G. C. Schatz, *J. Chem. Phys.*, 1996, **105**, 10472.
13 G. G. Maisuradze, D. L. Thompson, A. F. Wagner and M. Minkoff, *J. Chem. Phys.*, 2003, **119**, 10002.
14 Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner and M. Minkoff, *J. Chem. Phys.*, 2004, **121**, 5091.
15 X. Huang, B. J. Braams, S. Carter and J. M. Bowman, *J. Am. Chem. Soc.*, 2004, **126**, 5042–5043.
16 X. Zhang, B. J. Braams and J. M. Bowman, *J. Chem. Phys.*, 2006, **124**, 021104.
17 T. B. Blank, S. D. Brown, A. W. Calhoun and D. J. Doren, *J. Chem. Phys.*, 1995, **103**, 4129.
18 H. Gassner, M. Probst, A. Lauenstein and K. Hermansson, *J. Phys. Chem. A*, 1998, **102**, 4596–4605.
19 S. Lorenz, A. Groß and M. Scheffler, *Chem. Phys. Lett.*, 2004, **395**, 210–215.
20 A. Brown, B. Braams, K. Christoffel, Z. Jin and J. Bowman, *J. Chem. Phys.*, 2003, **119**, 8790.
21 X. Huang, B. J. Braams and J. M. Bowman, *J. Chem. Phys.*, 2005, **122**, 44308.
22 S. Manzhos and T. Carrington, *J. Chem. Phys.*, 2006, **125**, 194105.
23 S. Manzhos, X. Wang, R. Dawes and T. Carrington, *J. Phys. Chem. A*, 2006, **110**, 5295–5304.
24 E. Sanville, A. Bholoa, R. Smith and S. D. Kenny, *J. Phys.: Condens. Matter*, 2008, **20**, 285219.
25 C. M. Handley and P. L. A. Popelier, *J. Phys. Chem. A*, 2010, **114**, 3371–3383.
26 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
27 P. Steinhardt, D. Nelson and M. Ronchetti, *Phys. Rev. B*, 1983, **28**, 784.
28 J. Duijneveldt and D. Frenkel, *J. Chem. Phys.*, 1992, **96**, 4655–4668.
29 E. R. Hernández and J. Íñiguez, *Phys. Rev. Lett.*, 2007, **98**, 55501.
30 A. van Blaaderen and P. Wiltzius, *Science*, 1995, **270**, 1177.
31 R. Kondor, *A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum*, 2007, arxiv preprint cs/0701127.
32 J. Behler, *J. Chem. Phys.*, 2011, **134**, 074106.
33 H. Weyl, *The Classical Groups*, Princeton University Press, 1946.
34 S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola and P. Baldi, *Bioinformatics*, 2005, **21**, i359–i368.
35 M. Valle and A. Oganov, *Acta Cryst.*, 2010, **A66**, 507–517.
36 D. A. Varshalovich, A. N. Moskalev and V. K. Khersonskii, *Quantum theory of angular momentum*, World Scientific, 1987.
37 J. P. K. Doye, M. A. Miller and D. J. Wales, *J. Chem. Phys.*, 1999, **110**, 6896.
38 L. B. Pártay, A. P. Bartók and G. Csányi, *J. Phys. Chem. B*, 2010, **114**, 10502–10512.
39 C. Chakravarty and R. M. Lynden-Bell, *J. Chem. Phys.*, 2000, **113**, 9239.
40 R. D. Mountain and A. C. Brown, *J. Chem. Phys.*, 1984, **80**, 2730.
41 R. Biswas and D. Hamann, *Phys. Rev. B*, 1987, **36**, 6434–6445.
42 R. Kakarala, *PhD thesis*, Department of Mathematics, UC Irvine, 1992.
43 H. Maschke, *Math. Ann.*, 1898, **50**, 492–498.
44 C. D. Taylor, *Phys. Rev. B*, 2009, **80**, 024104.
45 A. V. Meremianin, *J. Phys. A: Math. Gen.*, 2006, **39**, 3099–3112.
46 M. A. Caprio, K. D. Sviratcheva and A. E. McCoy, *J. Math. Phys.*, 2010, **51**, 093518.
47 N. Artrith and J. Behler, *Phys. Rev. B*, 2012, **85**, 045439.
48 D. Porezag, T. Frauenheim, T. Köhler, G. Seifert and R. Kaschner, *Phys. Rev. B*, 1995, **51**, 12947–12957.
49 J. Li, *Modell. Simul. Mater. Sci. Eng.*, 2003, **11**, 173.

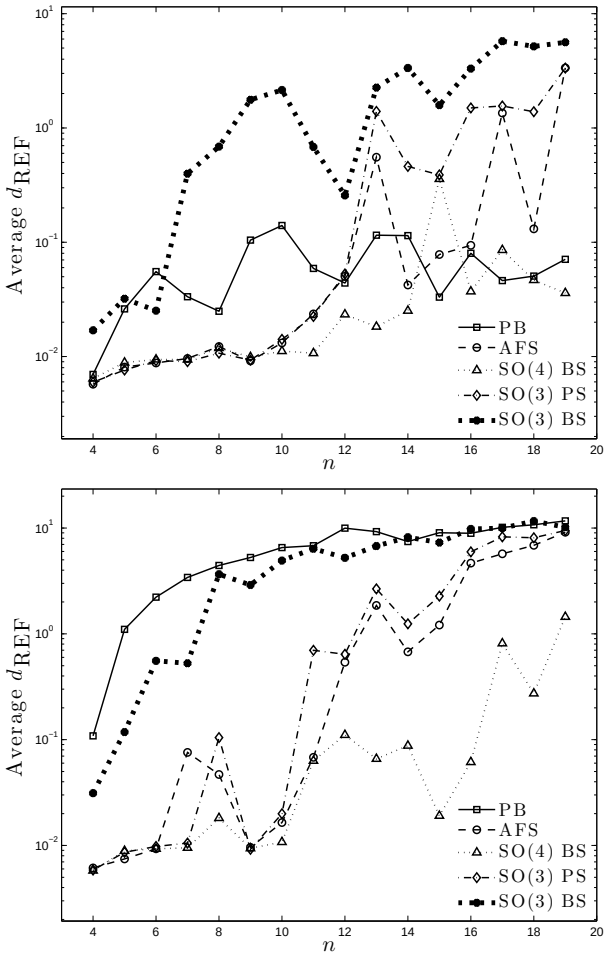

Fig. 7 Average difference after reconstructing structures randomised by 0.2 Å (top) and 1.6 Å (bottom) as a function of the number of atoms in the cluster. The cutoff was 6 Å. Different lines correspond to different descriptors: Parrinello-Behler (PS), Angular Fourier Series (AFS), bispectrum (BS) and power spectrum (PS). The two versions of the bispectrum differ in the handling of the radial degrees of freedom.

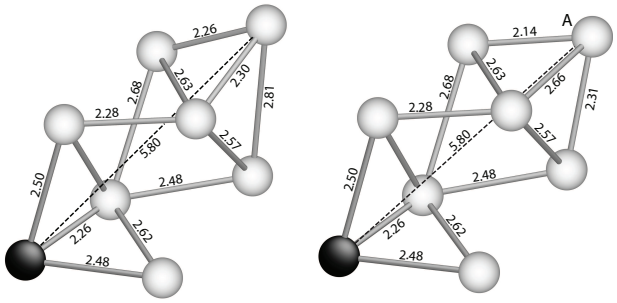

Fig. 8 Two Si_8 clusters that differ by $d_{\mathrm{ref}} = 0.7\text{Å}^2$. The reference atom, *i.e.* the centre of the rotation, is coloured black. The only difference between the two clusters is the relative position of the furthest atom, A.

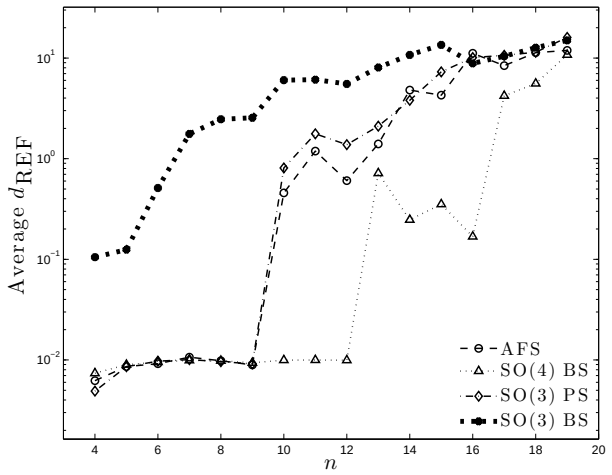

Fig. 9 Average difference after reconstructing structures randomised by 1.6 Å as a function of the number of atoms in the cluster. The cutoff was 9 Å. The line types corresponding to the different descriptors are the same as in Figure 7.

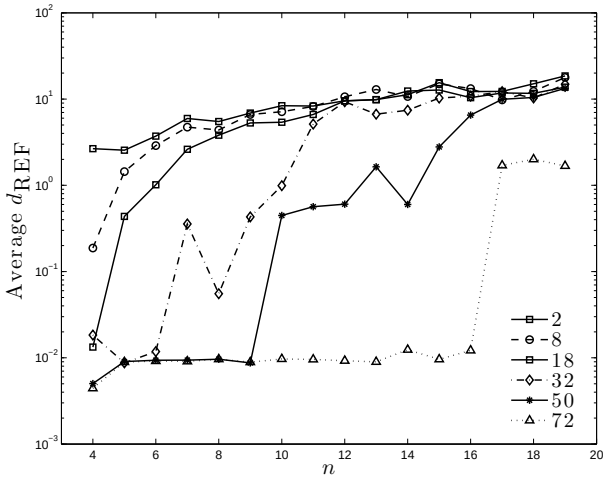

Fig. 10 Average difference after reconstructing structures randomised by 1.6 Å as a function of the number of atoms in the cluster using the AFS descriptor with a radial cutoff of 9 Å. The different curves correspond to different number of components of the AFS descriptor, achieved by varying the truncation of the angular expansion.

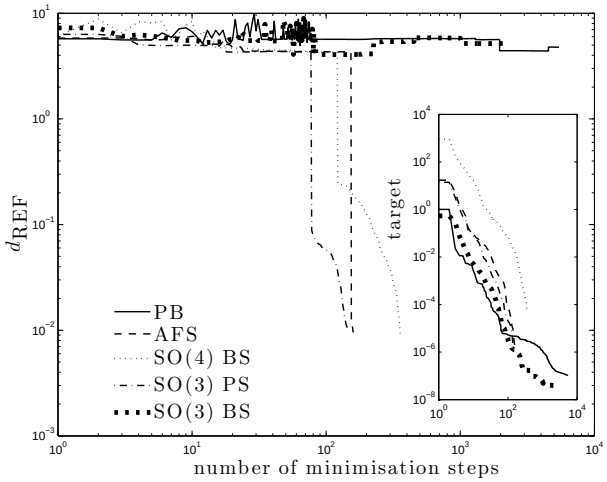

Fig. 11 Convergence of reference distance measure during the reconstruction procedure. The inset shows the value of the minimisation target approaching zero, i.e. descriptor equivalence.